EKONOMETRIA ECONOMETRICS 2(52) • 2016

ISSN 1507-3866
e-ISSN 2449-9994

Radosław Mącik
Maria Curie-Skłodowskia University
e-mail: radoslaw.macik@umcs.pl

VISUALISATION OF NOMINAL DATA – PRACTICAL AND THEORETICAL REMARKS

WIZUALIZACJA DANYCH MIERZONYCH NA SKALI NOMINALNEJ – UWAGI PRAKTYCZNE I TEORETYCZNE

DOI: 10.15611/ekt.2016.2.02
JEL Classification: Y10, C10, D12.

Summary: Nominal data, due to their nature, are often analysed statistically in a quite limited and traditional way. Usually they come from open-ended or simple/multiple choice questions. In typical research projects, such data are often presented in the form of more or less complex tables (including contingency tables) and standard charts. The author's experience shows that such a visualisation is perceived as boring, especially by younger people, accustomed to the presentation of content in the form of infographics. The article presents examples of data analysis and a visualisation of the nominal data based on the results of the author's research, including theoretical reflections on the techniques and tools used. The starting point is the raw text data from the responses to the open-ended questions subjected to analyses of the frequency of words and expressions, including its visualisation through word clouds. The next step is categorization and tabulation at the level of individual variables including the visualisation of categories, to assess the contingency between two nominal variables (or the nominal and the ordinal one), including visualising the relationships via chord diagrams and the correspondence analysis.

Keywords: Text data, nominal data, visualisation, word cloud, word tree, chord graph, correspondence analysis.

Streszczenie: Dane nominalne ze względu na swój charakter często są analizowane statystycznie w dość ograniczony i na ogół tradycyjny sposób. Zazwyczaj pochodzą one z pytań otwartych albo pytań o strukturze prostej lub wielokrotnej kafeterii. W typowych projektach badawczych takie dane są prezentowane zazwyczaj w formie mniej lub bardziej rozbudowanych tabel (w tym tabel kontyngencji) oraz standardowych wykresów. Z doświadczenia autora wynika, że taki sposób wizualizacji wyników odbierany jest jako nudny, szczególnie przez młodsze osoby, przyzwyczajone do prezentacji treści w formie infografik. Artykuł prezentuje przykłady analizy i wizualizacji danych o charakterze nominalnym na podstawie wyników własnych badań autora łącznie z refleksją teoretyczną na temat stosowanych technik i narzędzi. Punktem wyjścia są surowe dane tekstowe pochodzące z odpowiedzi na pytania otwarte, które zostały poddane analizom częstości występowania słów i wyrażeń, łącznie z jej wizualizacją za pomocą chmur słów. Kolejny prezentowany etap to kategoryzacja i tabulacja na

poziomie pojedynczych zmiennych łącznie z wizualizacją kategorii. Zasadniczy obszar prezentowanych analiz dotyczy kontyngencji zazwyczaj dwu zmiennych nominalnych (lub jednej nominalnej, a drugiej porządkowej) i sposobów jej wizualizacji za pomocą m.in. wykresów strunowych i technik analizy korespondencji.

Słowa kluczowe: dane tekstowe, dane nominalne, wizualizacja, chmura słów, drzewo słów, diagram strunowy, analiza korespondencji.

1. Introduction

In marketing research, quantitative and qualitative approaches are commonly used. The former concentrates on measurement via scales in questionnaires or mechanical/physiological instruments, and measurement results are interpreted as numbers typically treated with statistical analysis. The latter approach is interpretative in its nature and research material comes usually as natural language text and images; drawings as well as projective techniques results from individual or group in-depth interviews. In both approaches, the data on nominal measurement level are typical, in questionnaire research coming from open-ended questions and simple/multiple choice question structures. Transcriptions of qualitative interviews and open-ended question responses provide textual material and are typically coded into multiple categories, usually creating a nominal or ordinal variable. Only the nominal measurement level is the point of interest in this paper.

Nominal (or categorical) variables are variables whose values do not have a natural ordering [Rosario et al. 2004], so it is only possible to count the occurrences of particular categories regarding frequencies. Coding of the text often leads to a large number of categories (distinct values), and these variables are often called the high cardinality nominal variables [Rosario et al. 2004]. Such variables are common in real-world data sets, not only coming from survey data in social science. Typical examples of high cardinality nominal variables include product names and codes, species names in biology, and country or other location names in demographical descriptions of the study participants.

The scope of the nominal variables statistical treatment is somewhat limited, particularly to the tabulating and visualisation of categories, as well as cross-tabulation with chi-square tests and correspondence analysis as the most common techniques used in this case. For high cardinality nominal variables, another set of problems arises:

1. How to visualize, for instance, 30-50 distinct categories' frequencies? (When typical graphs are in this case unreadable?).

2. How to assess and visualize connections between these categories? (When in cross tabulation we have tables with hundreds of cells, mostly with zero frequencies?).

One option is to aggregate several categories into a few broader ones, but this means effectively significant information loss. Another possibility is to use less

conventional approaches such as word clouds and word trees to visualize text responses, and for more aggregated data – particularly for visualizing relationships – techniques like chord diagrams and correspondence analysis.

The need to focus on visualisation arises with changes in information processing habits by information users. In particular younger audience is accustomed to graphical, and often interactive, presentation of data, particularly in the form of infographics increasing its popularity. So the traditional presentation of data in the form of tables and classical graphs is often perceived as annoying and not attractive.

From another point of view, data visualisation enhances cognition, particularly by helping to think in a suggested, specific way [Card et al. 2009]. Proper data visualisation amplifies cognitive processes by: *"(a) increasing the memory and processing resources available to the users, (b) reducing the search for information, (c) using visual representation to enhance the detection of patterns, (d) enabling perceptual inference operations, (e) using perceptual attention mechanisms for monitoring, and (f) by encoding information in a manipulable medium.*" [Card et al. 2009, p. 187]. The feature mentioned last applies to dynamic diagrams, so it is beyond the scope of this paper.

The primary goal of this article is to present the examples of data analysis and a visualisation of the nominal data based on the results of the author's research, including theoretical reflections on the techniques and tools used, under the assumption of free and possibly open-source tools usage. Also, simple to use tools were selected, avoiding the usage of R-packages and data mining software. The used approach assumes the possibility to carry-out described visualisations by a researcher with limited knowledge in data analysis techniques. The starting point is the raw text data from the responses to the open-ended questions subjected to analyses of the frequency of words and expressions, including its visualisation through word clouds. The next step is the categorization and tabulation at the level of individual variables. After the visualisation of categories is made to assess the contingency between two nominal variables (or the nominal and the ordinal one), including visualising the relationships via chord diagrams and the correspondence analysis (only one carried out in SPSS environment, although free tools for CA are also available).

2. Methodological note

Data used as examples throughout this paper come from open-ended questions included in the CAWI questionnaire. Participants were expected to choose, using one of the two most commonly used internet price comparison engines, an automatic coffee machine as a courtesy purchase suggestion for a neighbor who does not use the internet. After this task, they were asked about the choice made and the primary reasons to choose a particular product, and its seller in the virtual sales channel using a set of open-ended questions. The research was conducted during March 2015 through the CAWI questionnaire with an e-mail invitation sent to the author's

students and their peers, that returned 461 usable responses from 575 sent messages, giving a response rate of 80.2%. Students received a small increase in course activity grade in exchange for the participation and recruitment of their peers (this award was less than 4% of the total possible grade).

The sample consisted of 60.2% women and 39.8% men. The average age of the participants was 24.5 years with a standard deviation of 5.1 years (range: 18-46 years old, median: 23 years) and is diversified regarding the place of residence – each third of the participants were from rural areas, small towns and larger cities. All the participants must be active internet users and make at least one online purchase during the year before the study. The sample structure regarding gender and age resembles the population of full-time and part-time students of a public university located in south-East of Poland, where the data were collected.

The presented examples come from two sets of text responses: reasons to choose the particular product (Text 1: 913 expressions consisting of 3986 words total and 693 different words) and reasons to choose the seller (Text 2: 901 expressions with 3836 words total and 638 different words). As the research was carried out in the Polish language, both texts were translated into English and back-translated by a native speaker to provide equivalence of both versions.

3. Raw text visualisations

3.1. Word clouds

Word clouds have become in recent years one of the most widely accepted techniques for the visualisation of keywords extracted from textual data [Paulovich et al. 2012]. A word cloud (also called a tag cloud) is quite sophisticated, but for the user easy-to-interpret method of summarizing and visualizing the content of a large quantity of text, by depicting the most common terms appearing within the text provided via font size relative to the frequency of each word in the text. In other words "*word cloud is a special visualization of text in which the more frequently used words are effectively highlighted by occupying more prominence in the representation*" [McNaught, Lam 2010, p. 630].

Word clouds can be built directly from a text using the word frequencies, after eliminating the stop words. Stop words are typically a collection of words from a specified language that do not contain relevant information, but rather are used to connect other words grammatically. The words in a word cloud are often sorted in alphabetical order, which provides no information, although it helps to retrieve information contained in the cloud, by easier finding a particular word, particularly a less common one [Gambette, Véronis 2010]. Existing methods are adequate to demonstrate content but are not capable of preserving much of the semantic relationships among keywords [Paulovich et al. 2012], which is an evident drawback of word clouds.

Word clouds were used previously for analyzing pieces of literature and political, as well as for extracting the most common words from transcribed focus group interviews (also by the author). Before visualisation is possible, there is the need for preprocessing the analysed text (externally or internally) within the used tool. Preprocessing requires following these typical steps:

1. Exclude common words (stop words) in the language of the text, usually with the help of freely available files containing a list of such words.
2. Exclude other unwanted words from their own list (optional).
3. Group similar words by reducing to a stem – the stemming procedure (if available for the particular tool and language).
4. Optionally: change all the text to lowercase.

Typical parameters to decide what the word cloud will contain and look like include:

- The number of words to include – typically 25-150 depending on the total number of different words (it is reasonable to choose about 10-15% of them or the minimum frequency of words in data.
- Font and color set choices (mostly as a decorative feature).
- Formula for differentiating font size – most commonly as directly proportional to the frequency of word (n), scaled as square root from the frequency or the logarithm of frequency.
- The spatial arrangement of words – typically positioned in a spiral way from the centre of the cloud according to decreasing frequencies or alphabetically ordered from the upper left corner of the cloud, sometimes also fitting into a particular shape.

Figure 1 presents simple word clouds created without stemming, coming from the TagCrowd.com free tool, with frequencies of words incorporated into visualisation.

a) Cloud from Text 1 – reason to choose product b) Cloud from Text 2 – reason to choose seller

Figure 1. Simple word clouds with frequencies of words

Source: own elaboration for the 70 most common words using the TagCrowd.com tool.

From the cloud based on Text 1, one can easily find that the most often used criteria to choose an automatic coffee machine were price, brand and meeting (other) requirements, as well as (positive) reviews. For Text 2, the highest frequencies of occurrence were for words such as price, store, free, lowest, free, shipping, suggesting the domination of price-related factors in choosing the seller: lowest price, free delivery/shipping.

a) Cloud from Text 1 – font size proportional to frequency

b) Cloud from Text 1 – font size as square root of word frequency

c) Cloud from Text 1 – font size as logarithm of word frequency

Note: same scaling factor used for all the word clouds, the tool uses a different set of stop words compared to TagCrowd.com.

Figure 2. Visualisation effect of function deployed on word frequency on word cloud perception

Source: own elaboration for the 70 most common words using the https://www.jasondavies.com/word-cloud/ tool.

The next example – Figure 2 – shows differences in the perception of word clouds using different functions for representing frequencies of words regarding font size in the case of the most common words from Text 1.

In part a) of Figure 2, there is a visible domination of the word "price" compared to other words in the cloud which, taking into account the word counts (222 for "price", 96 for "brand" and 82 for "meets"), looks adequate and does not create misunderstanding. This tool in some cases has problems with word alignment when n is scale factor, also the word "good" treated as a stop word by the TagCrowd.com tool used in Figure 1. Comparing part a) of Figure 2 with parts b) and c), when other scaling factors are used, it becomes evident that using log n (part c) made the word cloud almost useless. Words with frequencies differing by factor 3 or 4 are looking in this case similar, and using sqrt n is reasonable – the cloud is readable and the differences visible enough.

a) Cloud from Text 1 – reason to choose the product

b) Cloud from Text 2 – reason to choose the seller

Note: same scaling factor used for both word clouds, the tool uses its own set of stop words

Figure 3. Word clouds created with the Wordle tool

Source: own elaboration for the 70 most common words using the Wordle.net tool.

The commonly used Wordle tool [Feinberg 2014; McNaught, Lam 2010], provides aesthetically pleasing word clouds with the frequency of words used as the font scaling factor. It includes options to exclude stop words for 30 languages automatically, change the case of letters and provides many options to customize the created cloud in terms of font, arrangement of words and color palette.

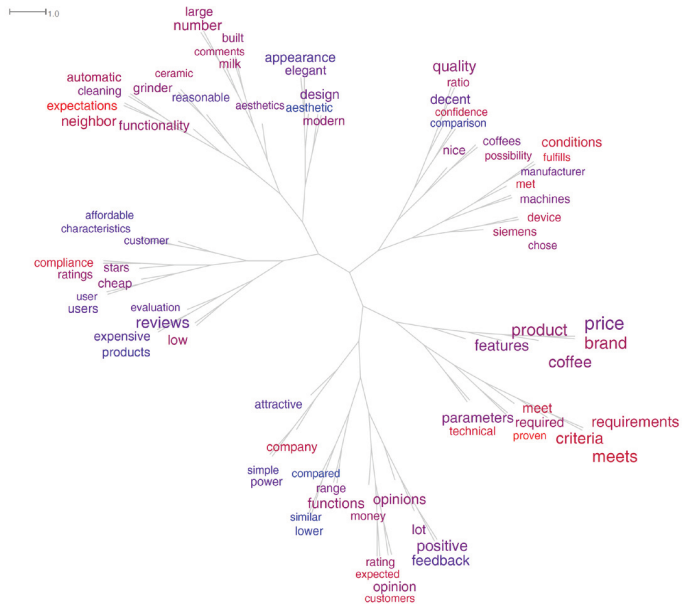
Examples of word clouds created using the Wordle.net tool are presented in Figure 3. Comparing visualisations with previously described tools word clouds from Wordle is much easier to interpret in the opinions of users, as stated during short interviews conducted by the author.

All the presented tools in a better or worse way visualize separate questions on the level of individual words, but there is no possibility with them to analyse word collocations (which words are close to each other in the responses) and also to directly compare data from two separate questions (in this case, for example, check if price as the criterion to choose the product was also the criterion to choose the seller by particular respondents.

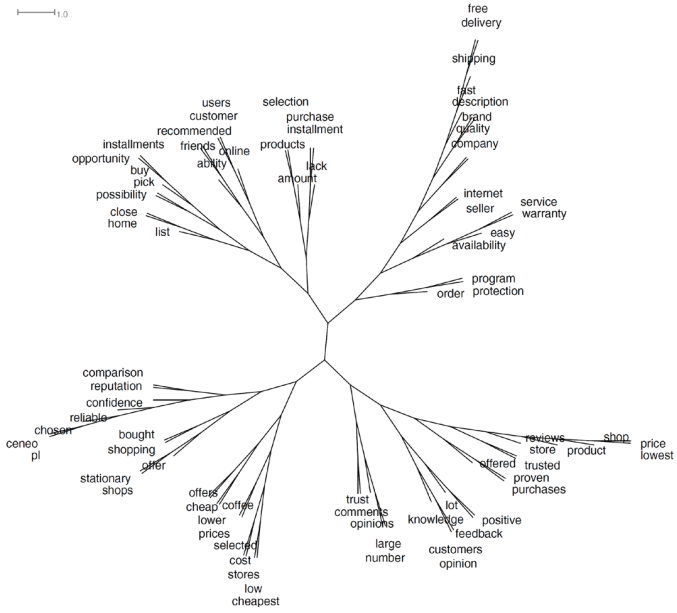
3.2. Word trees

A more advanced possibility than word clouds is to create from the analysed text so called word trees [Gambette, Véronis 2010]. A word tree, similarly to a word cloud, shows the most frequent words of the text, where the font size for particular word reflects its frequency, but the words are arranged on a tree to reflect their semantic proximity on the basis of the analysed text. The concept to use a tree to reflect the semantic distance between words of a tag cloud comes from the work of Gambette and Véronis [2010]. In this approach "*the distance between two words is given by the length of the path between them in the tree*" [Gambette, Véronis 2010, p. 562]. As Gambette and Véronis explain "*the problem of finding a tree which reflects a distance matrix was introduced in bioinformatics to reconstruct phylogenetic trees from the information on the distances between their leaves*" [Gambette, Véronis 2010, p. 562]. Bioinformatics has provided algorithms that proved to be useful also for text processing to represent the proximity of words. The above mentioned authors wrote a free tool called TreeCloud that uses Python scripts and SplitsTree software from the bioinformatics field to produce word trees.

The TreeCloud tool was used to visualise both analysed sets of responses (Text 1 and Text 2, limiting the number of words to the 70 most common ones in both cases, exactly as for word clouds. The obtained results were shown in Figure 4 for reasons to choose the product (part a) and reasons to choose the seller (part b). Part a) is visualized with all the features the TreeCloud provides, including changing font size depending on word frequency as well as coloring the words. For comparison in part b) there is only a simple word tree without additional formatting shown. Obviously, supplementary information about word frequency adds much to the scope of the information possible to extract from the visualisation, so the form presented in Figure 4, part a) is more useful, the latter (part b) informing only about word collocations.



a) Word tree from Text 1 (reason to choose product) – full software features applied



b) Word tree from Text 2 (reason to choose seller) – basic formatting used

Note: same scaling factor used for both word trees.

Figure 4. Word trees examples

Source: own elaboration for the 70 most common words using the TreeCloud tool.

In both cases, visible branches of the tree are creating cluster-like groups depicting the most common collocations, for instance “product price” and “product brand” (part a). Also for word trees, there is no direct possibility to connect answers from both texts, and categorization is still needed.

4. Visualisation of categories contingency

The contingency of the two nominal variables is often hard to visualise. The standard for such a presentation is a traditional contingency table with appropriate chi-square tests or correspondence analysis.

In the provided example, data from Text 1 and Text 2 were coded separately, firstly into eight and nine categories respectively. Next, some categories were grouped to 5 for each variable. This information is presented in Table 1, suggesting that the declared factors to choose are independent of the stated factors to choose the store (merchant) (chi-square=14.155; p=0.587, df=16), and further implying that both decisions are separate for most respondents that leads to a change of thinking about the phases of the consumer decision-making process in an ICT mediated shopping environment.

Table 1. Contingency between reasons to choose product vs. store

Main store choice reason:	Main product choice reason:					
	Low, attractive price (prod_ price)	Known, trusted brand (prod_ brand)	Attractive design (prod_ design)	Functions, features (prod_ features)	Good opinions, reviews (prod_ reviews)	Total
Lower price, free delivery (store_price)	61.2%	51.3%	44.9%	60.5%	50.0%	55.3%
Known, trusted store brand (store_brand)	13.2%	22.5%	24.5%	11.1%	17.0%	16.4%
Favorable opinions (store_ reviews)	13.2%	16.3%	16.3%	12.3%	20.5%	15.3%
Offer positioning, web creation (store_positioning)	5.3%	2.5%	6.1%	4.9%	5.7%	4.9%
Availability of product in the store (store_availability)	7.2%	7.5%	8.2%	11.1%	6.8%	8.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Note: in parentheses are denoted short descriptions of categories used in Figure 5 and Figure 6.

Source: own research.

The structures of responses in the columns are similar, and the chi-square independence test suggest rejecting the hypothesis about the dependence of the main reason to choose an automatic coffee machine with the main reason to choose the

store selling it. In particular the frequency of declaring the lowest price and free delivery in a price sensitive group is not significantly different than in other groups.

4.1. Chord diagrams

The traditional contingency table, easy to interpret by a reading-oriented audience, is not so easy understandable for visually-oriented persons, particularly young individuals seeking easy, simple visual stimuli. For such users, so called chord diagrams can help in understanding the patterns in the data.

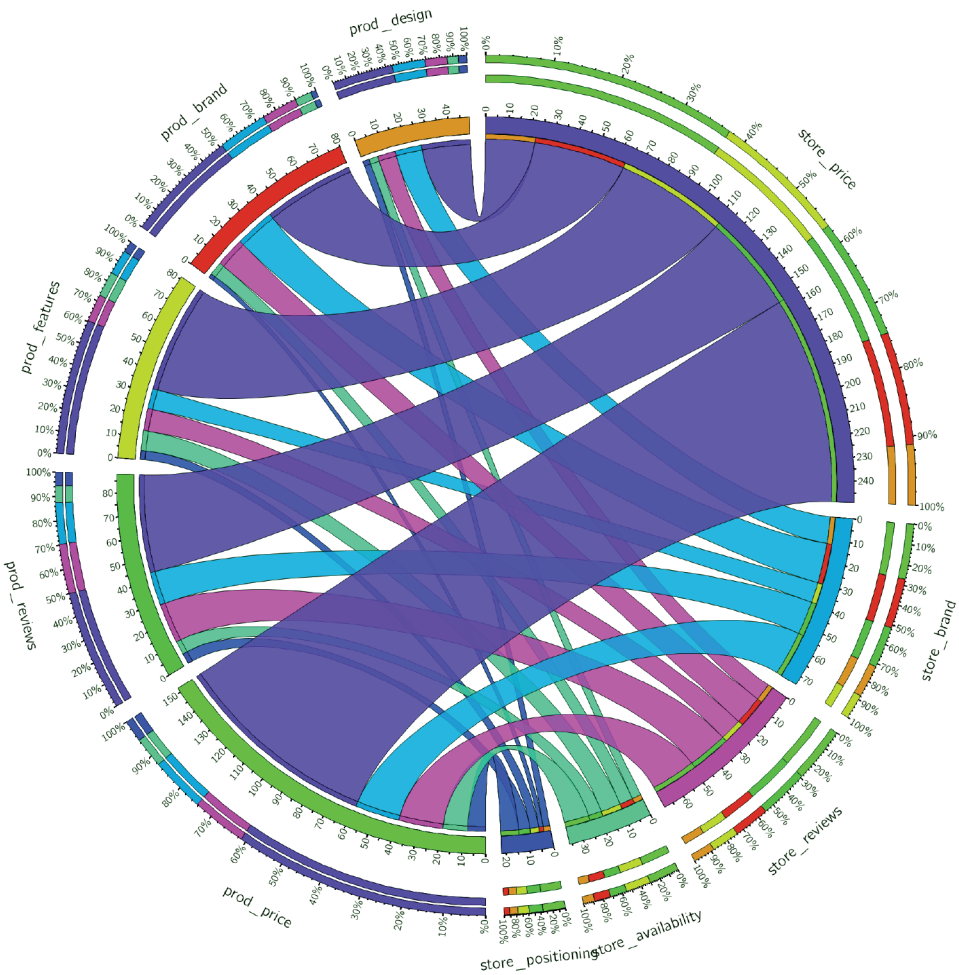


Figure 5. Contingency visualisation via chord diagram

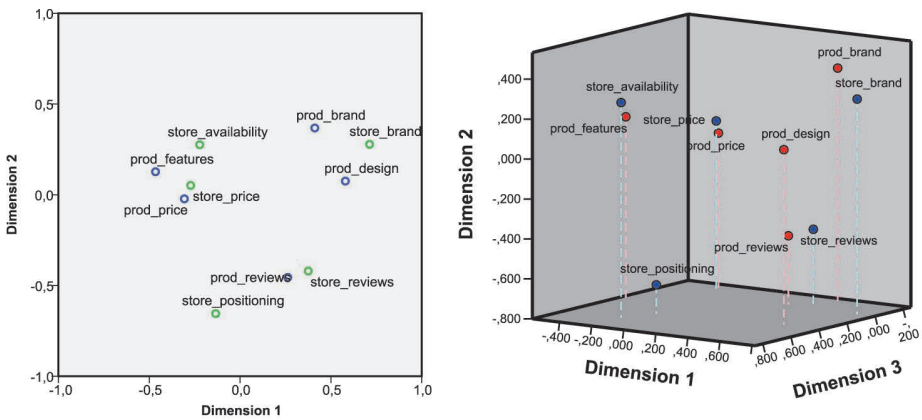
Source: own elaboration with the Circos Table Viewer tool (http://mkweb.bcgsc.ca/tableviewer/visualize/).

Chord diagrams show directed relationships among a group of entities, like migration visualisation [Abel, Sander 2014]. Easy to use tools for the creation of chord diagrams like Circos can be effectively used to represent tabular data graphically, although it is a more powerful tool used mainly in genomic research [Krzywinski et al. 2009]. In this application, the ideograms "*represent individual rows or columns of a table, and a ribbon* […] *represents the value of a cell for a given row and column*" [Krzywinski et al. 2009, p. 1644].

The chords on right side of the diagram (Figure 5) exhibit a similar pattern for each group of reasons to choose a product, in particular the ribbon for the tendency to seek a low price including free delivery is the thickest of all the chords, confirming the lack of dependency between the factors of product and store choice.

4.2. Correspondence analysis

In the case of seeking a visual representation of contingency of nominal variables, a simple or multiple correspondence analysis (CA and MCA) can be useful. Particularly the graphical representation of CA results can help to find similarities between row and column categories, although sometimes it is better to analyse performance in three dimensions than in two (when it is significant regarding CA solution quality – this is beyond the scope of this paper). Figure 6 presents graphs obtained from a simple correspondence analysis carried in SPSS package in 2D (part a) and 3D (part b).



a) CA results visualized in two main dimensions b) CA results visualized in three main dimensions

Figure 6. Correspondence analysis results visualisations

Source: own elaboration with SPSS, part b) produced using GPL syntax from point coordinates.

The provided example shows that 2D visualisation suggests a greater similarity between row and columns points than 3D – introducing an additional dimension

breaks visual clusters, differentiating for example between product design and product/store brand as a factor of choice. Similarly, store positioning and website features start to differ from the product and sellers reviews. It is worth to note that the graphical representation of (dis)similarities in CA visual output should not be over-interpreted, particularly when CA measures of fit are far from perfect, including the not significant results of chi-square independence tests.

5. Conclusion

The examples shown in this paper and literature on the subject of the visualisation of raw text data are suggesting the lack of need to code the analysed text at the early stages of analysis. At these stages, tools like word clouds and word trees can be used to extract the more frequently used words and visualize connections between them, mostly regarding spatial or semantic similarity. The same applies to high cardinality data – word clouds visually enhance the main content. The proper construction of such visualisations allows the user to perceive the most relevant information "at first sight," although there is no possibility to analyse the contingency between different variables.

Coding into a few categories can be done later, and there is no need to lose information early. From the shown examples one can conclude that chord diagrams and graphs from the correspondence analysis can be visually appealing and easier to interpret compared to a traditional table.

Tools for the aesthetical and methodologically correct visualisation of nominal data are nowadays often freely available and straightforward to use. Proving that was also one of the goals of this paper.

Bibliography

Abel G.J., Sander N., 2014, *Quantifying global international migration flows*, Science, 343(6178), pp. 1520–1522.

Card S., Mackinlay J.D., Shneiderman B., 2009, *Information Visualization. Human-Computer Interaction: Design Issues, Solutions, and Applications*, CRC Press, Boca Raton, London, New York.

Feinberg J., 2014, *Wordle*, Available at: http://www.wordle.net/ [Accessed October 6, 2015].

Gambette P., Véronis J., 2010, *Visualising a Text with a Tree Cloud*, [in:] H. Locarek-Junge, C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin Heidelberg, pp. 561–569.

Krzywinski M. et al., 2009, *Circos: An information aesthetic for comparative genomics*, *Genome Research*, 19(9), pp. 1639–1645.

McNaught C., Lam P., 2010, *Using Wordle as a supplementary research tool*, The Qualitative Report, 15(3), p. 630.

Paulovich F.V. et al., 2012, *Semantic Wordification of Document Collections*, [in:] *Computer Graphics Forum*, Wiley Online Library, pp. 1145–1153. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8659.2012.03107.x/full [Accessed June 13, 2016].

Rosario G.E. et al., 2004, *Mapping nominal values to numbers for effective visualization,* Information Visualization, 3(2), pp. 80–95.